A Fresh Look at Multicanonical Monte Carlo from a Telecom Perspective

Alberto Bononi* , Leslie A. Rusch†, Amirhossein Ghazisaeidi†, Francesco Vacondio† and Nicola Rossi*
* Dip. Ing. Inf., Università di Parma, 43100 Parma, Italy † ECE Dept. Université Laval, Québec, Québec, Canada G1K 7P4
Email:alberto.bononi@unipr.it, Email:rusch@gel.ulaval.ca

***Abstract*—The Multicanonical Monte Carlo (MMC) technique is a new form of adaptive importance sampling (IS). Thanks to its blind adaptation algorithm, it does not require an in-depth system knowledge for exploitation as does traditional IS. Hence MMC is a practical, handy tool to estimate via simulation the probability of rare events in complex telecom systems, such as the symbol error rate or the outage probability. In this paper, we present the analytical connections between MMC and IS, and describe the recursive algorithm via which MMC seeks an optimal "flat-histogram" warping. We also provide practical guidelines on how MMC can be successfully applied in telecom to achieve accelerations of simulation time by many orders of magnitude with respect to standard Monte Carlo.**

***Index Terms*—Simulation, Monte Carlo methods, importance sampling, adaptive algorithms**

I. Introduction

THE Multicanonical Monte Carlo (MMC) technique, developed by physicists Berg and Neuhaus fifteen years ago [1], offers a new powerful tool to telecom engineers for simulation of systems which are difficult to attack by analytical or semi-analytical means.

Berg's papers are hard to read for non-physicists, and the probability theory ideas are hidden by a plethora of statistical physics details. For this reason, the first telecom community in which MMC was used was optical communications, where physicists and electrical engineers share a common background and a common language. Physicists Yevick and Menyuk [2], [3] first applied MMC to optical communications to find the bit error rate (BER) and the outage probability due to fiber polarization mode dispersion. Soon after those publications, a large number of MMC papers appeared on various topics in optical communications [4]–[17].

The success of MMC is mostly due to its ease of implementation. MMC is a method to estimate the entire distribution of a desired system output variable Y (a scalar in its simplest form and as described in this paper), given the known distribution of the system multi-dimensional input X. It is related to the well known method of importance sampling (IS) [18]. The most striking shortcoming of traditional IS is that an in-depth knowledge of the physical problem at hand is required to find an efficient biasing distribution, making IS time-consuming in its planning phase and thus difficult to use. MMC is instead a truly innovative adaptive IS algorithm, which automatically searches the optimal biasing distribution for the estimation of the entire distribution of Y, namely, the biasing distribution providing a flat output histogram. Although it has been shown that IS can be slightly more efficient than MMC in the estimation of the probability of rare events [7], MMC has the big advantage of being easily implemented for any system, with great time savings in the planning phase.

The key tool used by MMC to adaptively generate biased distributions with a desired density is the Markov Chain Monte Carlo (MCMC) method [19], [20], which is routinely used in statistical physics to sample from Boltzmann-like input distributions [21]. To the authors' knowledge, all published papers on MMC so far delve into the machinery of the MCMC method, as if the true heart of the MMC algorithm were the MCMC biasing scheme. In this paper, we instead explain MMC without the need of MCMC, so that all attention can be focused on the explicit analytical connections between MMC and IS. Later MCMC will enter into play, but its function within MMC will be clear and the reader will better appreciate the subtleties connected with its use within MMC.

This paper is organized as follows. After a brief review of classical Monte Carlo (MC) in Sec. II-A, importance sampling (IS) is introduced in Sec. II-B with a new twist with respect to classical treatments [18]; the concepts of uniform weight (UW) IS and flat histogram (FH) IS are introduced. Then the MMC FH adaptation algorithm is described in Sec. III-A, and practical aspects of MMC are discussed in Sec. IV. The Appendix contains a summary of MCMC.

II. Monte Carlo Techniques

In order to determine the symbol error rate (SER) of a digital communications system we need the statistical properties of the decision variable at the output of the receiver. Let that decision variable be $Y = g(X)$, where $g : \Gamma \to \mathbb{R}$ is a real scalar function of a random vector X taking values in the *input* (or *state*) *space* Γ. We are interested in determining the distribution (i.e., the probability density function (PDF) in the continuous case or the probability mass function (PMF) in the discrete case) of Y. The system input-output transfer function $g(\cdot)$ is in most practical problems known only through a computationally expensive numerical routine. We assume the joint PDF $f_X(x)$ of X (or equivalently the joint PMF in the discrete case) is known, possibly up to an unknown multiplicative constant, and we are able to draw samples from such a distribution.

In digital communications, the system random input X is the set of all noise samples accumulated along the transmission line, and all the random symbols in the transmitted sequence, that fall within a memory window around the sampling instant and together determine the random value of the decision variable Y. The larger the memory of the transmission system, the larger the dimensionality of X. In the rest of this paper, we will assume that Y and X are continuous random variables (RVs). The modifications for discrete RVs are straightforward.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

A. Conventional Monte Carlo Estimation

In order to estimate by simulation the PDF $f_Y(y)$ of the continuous output Y on a desired range $\mathcal{R}_Y$, we tile $\mathcal{R}_Y$ with M bins of width Δy centered at the discrete values $\{y_1, ..., y_M\}$. [1] We define the i-th bin as the interval $B_i \triangleq \left[y_i - \frac{\Delta y}{2}, y_i + \frac{\Delta y}{2}\right]$. If the PMF of the discretized Y on the i-th bin is $P_i \triangleq P\{Y \in B_i\}$, then for sufficiently small Δy the output PDF is $f_Y(y_i) \simeq P_i/\Delta y$. This binning implicitly defines, via $g(\cdot)$, a partition of the input space into M domains $\{D_i\}_{i=1}^M$, where

$$D_i = \{x \in \Gamma : \, g(x) \in B_i\}$$

is the domain in Γ that maps into the i-th bin. While B_i are simple intervals, the domains D_i are multidimensional regions with possibly tortuous topologies, and most often totally unknown to the researcher. Let the Bernoulli RV

$$I_{D_i}(X) = \begin{cases} 1 & \text{if } X \in D_i \\ 0 & \text{else} \end{cases}$$

be the indicator of event $\{X \in D_i\}$, or equivalently $\{Y = g(X) \in B_i\}$, which more clearly indicates that calculation of $g(X)$ is needed to determine if this event occurs. Then the desired PMF can be expressed as the expectation of the indicator

$$P_i = \int_{D_i} f_X(x)\mathrm{d}x = \int_\Gamma I_{D_i}(X) f_X(x)\mathrm{d}x = E[I_{D_i}(X)]. \tag{1}$$

This is the rationale behind classical MC estimation: draw N samples $\{X_1, .., X_N\}$ from the distribution $f_X(x)$, pass them through the system $g(\cdot)$ and find how these samples fall in the output bins, forming the histogram. The (normalized) histogram is the sample mean of the expectation of the indicator in (1), thus the following estimate of the PMF

$$\hat{P}_i^{MC} \triangleq \frac{1}{N}\sum_{j=1}^{N} I_{D_i}(X_j) = \frac{N_i}{N} \tag{2}$$

N_i being the number of samples that fall in bin i. The MC estimator is unbiased by construction: $E[\hat{P}_i^{MC}] = P_i$. The squared relative error (SRE), a figure of merit for any unbiased estimator $\hat{P}_i$, is defined as $\varepsilon_i \triangleq Var[\hat{P}_i]/P_i^2$. If the samples are independent, N_i is the sum of N independent Bernoulli RVs with "success" probability P_i and thus has a binomial distribution, i.e., $N_i \sim Binomial(N, P_i)$, so that the SRE for the MC estimator is

$$\varepsilon_i^{MC} = \frac{1 - P_i}{NP_i} \tag{3}$$

which is, for small P_i, approximately the inverse of the expected value $E[N_i] = NP_i$. For instance, about 100 counts on average are required to achieve a relative error $\sqrt{\varepsilon_i}$ of 10% in the estimation of P_i. However, in MC simulations most samples fall in the modal bins, and little or no samples fall in the area in which we are most interested, the tails of the PMF, thus dramatically increasing the relative error in the tails.

[1] If the output range $\mathcal{R}_Y$ is not the entire output space, $f_Y(y)$ will actually denote the conditional PDF $f_Y(y|Y \in \mathcal{R}_Y)$.

B. Importance Sampling

In order to reliably estimate the output PMF even in the tail bins (rare events), we artificially increase the number of samples falling in such bins using IS [18]. We re-write (1) as

$$P_i = \int_\Gamma I_{D_i}(x)\left[\frac{f_X(x)}{f_X^*(x)}\right] f_X^*(x)\mathrm{d}x = E^*[I_{D_i}(X)w(X)] \tag{4}$$

where $f_X^*(x)$, strictly positive for all x at which $f_X(x) > 0$, is a *warped* PDF of X, and $w(x) \triangleq f_X(x)/f_X^*(x)$ is the *IS weight*; E^* indicates expectation with respect to the distribution $f_X^*(x)$. The output PMF in the warped space is given by

$$P_i^* = \int_\Gamma I_{D_i}(x) f_X^*(x)\mathrm{d}x = E^*[I_{D_i}(X)].$$

The weighting function $w(x)$ plays an important role in generating the IS estimate of the unwarped PMF. To see this, consider the conditional density $f_X^*(x \,|\, X \in D_i) = \frac{I_{D_i}(x) f_X^*(x)}{P_i^*}$ and use it to rewrite P_i in (4) as

$$P_i = P_i^* \int_\Gamma I_{D_i}(x) w(x) \frac{f_X^*(x)}{P_i^*}\mathrm{d}x = P_i^* E^*[w(X) \,|\, X \in D_i]. \tag{5}$$

The IS estimator replaces the product in (5) by the product of their sample averages in the warped system

$$\hat{P}_i^{IS} = \underbrace{\left(\frac{N_i^*}{N}\right)}_{\triangleq \hat{H}_i^*} \underbrace{\left[\frac{1}{N_i^*}\sum_{n=1}^{N_i^*} w(X_{j_n})\right]}_{\triangleq \bar{w}_i} \tag{6}$$

The IS estimation is performed as follows: a conventional MC simulation is run in the warped system, i.e., by drawing N samples *from the warped PDF* $f_X^*(x)$. The MC estimate in the warped system is determined by the N_i^* samples falling in bin i and forming the so-called histogram of visits $\hat{H}_i^*$ [21] in the warped system. Hence the IS estimate $\hat{P}_i^{IS} = \hat{H}_i^* \bar{w}_i$ comes naturally from the product of the MC estimate of P_i^* in the warped system, $\hat{H}_i^*$, and the estimate $\bar{w}_i$ of $E^*[w(X) \,|\, X \in D_i]$. The weights $\bar{w}_i$ of estimates P_i^* provide the inverse transformation to take us back into the unwarped system. The count N_i^* is on average much larger than in an unwarped MC sampling if we make $f_X^*(x) \gg f_X(x)$ over the domain D_i. Interestingly, we can equivalently write the IS estimator (6) as

$$\hat{P}_i^{IS} = \frac{1}{N}\sum_{j=1}^{N} I_{D_i}(X_j) w(X_j) \tag{7}$$

which is the traditional way of introducing IS as the sample average of the expectation in (4) [18].

To determine the accuracy of the IS estimate using (7), let $W_{ij} \triangleq I_{D_i}(X_j) w(X_j)$. From (4), $E^*[W_{ij}] = P_i$, and thus the IS estimator (6) is unbiased. To find its variance, observe that

$$\begin{aligned} E^*[W_{ij}^2] &= E^*[I_{D_i}(X_j) w^2(X_j)] \\ &= P_i^* \int_\Gamma I_{D_i}(x) w^2(x) \frac{f_X^*(x)}{P_i^*}\mathrm{d}x \\ &= P_i^* E^*[w^2(X) \,|\, X \in D_i], \end{aligned}$$

978-1-4244-4148-8/09/$25.00 ©2009

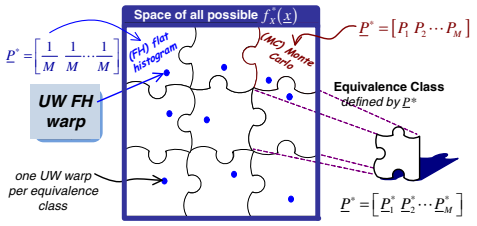

Figure 1. Sketch of the space of all input warpings $f_X^*(x)$, partitioned into disjoint equivalence classes, each characterized by a warped output PMF $\underline{P}^*$

so that from (7) we get

$$Var^*[\hat{P}_i^{IS}] = \frac{Var[W_{ij}]}{N} = \frac{P_i^* E^*[w^2(X)\,|\,X \in D_i] - P_i^2}{N}. \tag{8}$$

Using (5) the SRE $\varepsilon_i^{IS} \triangleq Var[\hat{P}_i^{IS}]/P_i^2$ becomes

$$\varepsilon_i^{IS} = \frac{1}{N}\left\{\frac{1}{P_i^*}\left(\frac{Var^*[w(X)\,|\,X \in D_i]}{(P_i/P_i^*)^2} + 1\right) - 1\right\}. \tag{9}$$

This formula helps us appreciate the true limit of IS estimation, which is connected to our *a priori ignorance of the domains* D_i. Suppose for instance that D_i is composed of two disjoint sets, located far apart on the input space: D_{i1} whose existence and location is found via physical reasoning and knowledge of our problem, and D_{i2}, of which we fail to guess the existence. This incomplete foreknowledge leads us to contrive a warping that shifts most of the PDF mass on D_{i1}, i.e., such that $f_X^*(x) \gg f_X(x)$, or equivalently we set $w(x) \ll 1$ on D_{i1}. Most likely we will get little PDF mass on D_{i2}, hence $f_X^*(x) \ll f(x)$, i.e. $w(x) \gg 1$ on D_{i2}, thus obtaining, as per (9), a very large value of $Var^*[w(X)\,|\,X \in D_i]$ and therefore a very large SRE.

C. Uniform Weight Importance Sampling

Now consider the set of all warpings $f_X^*(x)$ producing the same output warped PMF $\underline{P}^* \triangleq \{P_i^*\}_{i=1}^M$. We call this set the *equivalence class* of warpings associated with $\underline{P}^*$. The space of all possible warpings gets thereby partitioned into disjoint equivalence classes, as depicted in Fig. 1. From (5), each equivalence class produces the same average conditional weights $\{E^*[w(X)\,|\,X \in D_i]\}_{i=1}^M$. Equation (9) suggests that the best warping within each equivalence class, i.e., the one producing the lowest IS relative error, is the uniform weight (UW), which assigns a constant weight to all $x \in D_i$, with value $w_i = P_i/P_i^*$ as in (5), so that $Var^*[w(X)|X \in D_i] = 0$. *Hence, the search for the optimal global warping can always be restricted to the search among the UW warpings*. Note that although at first sight the implementation of UW warping seems to require a detailed knowledge of the domains D_i, we will shortly see that this is not the case.

From (9), the squared relative error for a UW-IS estimation of bin i simplifies to

$$\varepsilon_i^{UW-IS} = \frac{1}{N}\left\{\frac{1}{P_i^*} - 1\right\} \tag{10}$$

and depends only on P_i^*. When $P_i^* \ll 1$, the error is about the inverse of the expected value NP_i^*; this in turn is on average equal to the inverse of the warped count N_i^*. This leads to a reduced error with respect to ε_i^{MC} (3), at an equal number of runs N, on those bins in which the warping is doing well, i.e., in which $P_i^* \gg P_i$. In the extreme case when all warped samples fall in bin *i*, we reach the optimal UW-IS warping for estimating bin i. In this case $P_i^* \to 1$ and we achieve zero relative error; this is known as the *zero-variance* IS (ZV-IS) [18] warping. Such a warping will clearly be useless for the estimation of other bins.

Suppose we wish to use our N runs to estimate the output PMF on *all bins* with equally good relative error; (10) leads to the choice $P_i^* = \frac{1}{M}$ for all i. Since P_i^* is the expected value of the visits histogram, we will call this UW-IS the *uniform weight, flat-histogram (UW-FH)* importance sampling. It is easy to see that, among all UW-IS, the UW-FH is the one that minimizes the largest relative error among all bins, namely

$$\max_i \varepsilon_i^{UW-IS} = \max_i \frac{1}{N}\left\{\frac{1}{P_i^*} - 1\right\} \geq \varepsilon^{UW-FH} = \frac{M-1}{N}. \tag{11}$$

The analytic form of the warped input PDF $f_X^*(x)$ is needed, at least up to a normalization constant, to draw input samples from the warped system. Any UW warping can be expressed as [22], [23]:

$$f_X^*(x) = \frac{f_X(x)}{c_\Theta \Theta(x)} \tag{12}$$

where $\Theta(x) \triangleq \Theta_i$ for all $x \in D_i$, $i = 1 \ldots M$, and $\underline{\Theta} \triangleq \{\Theta_i\}_{i=1}^M$ is a positive PMF on the M bins (i.e., one with all non-zero entries), and c_Θ is a normalization constant. By construction, (12) puts constant weight $w_i = c_\Theta \Theta_i$ on each domain D_i.

The warped output PMF induced by such a UW warping is

$$P_i^* = \int_{D_i} f_X^*(x)\mathrm{d}x = \frac{\int_{D_i} f_X(x)\mathrm{d}x}{c_\Theta \Theta_i} = \frac{P_i}{c_\Theta \Theta_i}. \tag{13}$$

Since $\underline{\Theta}$ is by construction a proper PMF whose elements sum to one, the normalizing constant must be $c_\Theta = \sum_{j=1}^M \frac{P_j}{\Theta_j}$.

Consider next the particular UW-FH warping where $P_i^* = 1/M$. Equation (13) yields $c_\Theta = M$ and $\Theta_i \equiv P_i$. Hence from (12) the UW-FH warped PDF displays in its denominator the true PMF $\underline{P}$, which is exactly what we seek to estimate. Hence UW-FH appears unfeasible, like the ZV-IS, as it requires knowledge of exactly what we seek to estimate. We will show, however, that it can be closely approached by a sequence of UW warpings as in (12), via a simple adaptive mechanism.

III. Multicanonical Monte Carlo

Flat-histogram (FH) algorithms are a familiy of output PDF estimation algorithms, among which are MMC, Wang-Landau [24] and others [22]. Starting from the known input PDF $f_X(x)$, these algorithms build a sequence of UW-warped input PDFs $f_X^{(n)}(x) = \frac{f_X(x)}{c_n \Theta_n(x)}$, $n = 1, 2, ...$, in which the positive PMF $\underline{\Theta}_n \triangleq \{\Theta_{n,i}\}_{i=1}^M$ plays the role of an intermediate estimate of the true PMF $\underline{P}$ of the discretized output RV $Y = g(X)$ at the n-th step, and c_n is its normalizing constant. A step

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

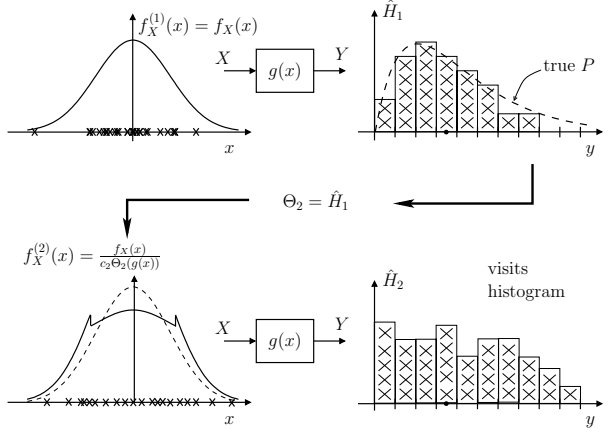

Figure 2. Sketch of first 2 steps in MMC. First cycle is a pure MC if we start with a uniform guess.

(which in MMC is called a *cycle*) corresponds to drawing of N samples $\{X_j\}_{j=1}^N$ from the warped $f_X^{(n)}(x)$, passing these samples through the system under test, and forming a new estimate $\underline{\Theta}_{n+1}$ of the PMF of Y. What characterizes a FH algorithm is its specific *update law* $\underline{\Theta}_n \to \underline{\Theta}_{n+1}$. In all cases, the update uses the output histogram of visits $\underline{\hat{H}}_n^* \triangleq \{\hat{H}_{n,i}^*\}_{i=1}^M$ at the end of cycle n, and drives this histogram in the next step towards equal visits to all bins (a flat histogram). At convergence, as seen from (12), $c_n \to M$ and $\underline{\Theta}_n \to \underline{P}$. Note that, no matter what the visits-flattening update law is, when the visits histogram is (practically) flat, the final estimate of the output PMF can be read off in the denominator of the warped input PDF, as we already noted at the end of the previous section.

A. MMC adaptation

MMC, introduced by Berg, *et al.*, in 1991 [1], is among the first FH methods. In MMC the update law is based on a UW-IS estimate. At cycle n, N samples are drawn from $f_X^{(n)}$ and $Y_j = g(X_j)$ is evaluated for every sample, finally forming the visits histogram $\hat{H}_{n,i}^* \triangleq N_{n,i}/N$. An IS-updated estimate of the PMF of discretized Y is obtained from (6) as

$$\Theta_{n+1,i} = \left(\frac{N_{n,i}}{N}\right)\left[\frac{1}{N_{n,i}}\sum_{n=1}^{N_{n,i}} w(X_n)\right] = \hat{H}_{n,i}^* c_n \Theta_{n,i} \quad (14)$$

where we used the constant weight $w_i = c_n\Theta_{n,i}$ of the previous warp $f_X^{(n)}$. In practice, c_n may be omitted, as per (26) in the appendix.

Figure 2 sketches the first two steps of MMC for the simple system $y = x^2$, with X a zero-mean Gaussian scalar RV. It is common practice to start the recursion (14) by using the uniform distribution as an initial guess for $\underline{\Theta}_1$, so that, as seen from (12), the first MMC cycle is performed with the unwarped distribution, i.e., as a classical MC run. In the example of Fig. 2, the bell-shaped input PDF $f_X^{(1)} = f_X$ is shown in the top left: most input samples (the crosses shown on the x axis) will fall on the modal region, and the output histogram will be an MC estimate of the true PMF, with a well-estimated modal region and almost no samples in the tails. At the end of the first cycle the PMF estimate (14) is updated to $\underline{\Theta}_2$ and used in the denominator of the warped input PDF at the next cycle. As sketched in the figure, the warped PDF $f_X^{(2)} = \frac{f_X}{c_2\Theta_2(x)}$ will decrease the mass function in the bins of the modal region in proportion to their number of visits, and increase the mass function in the tails after re-normalization by c_2. To avoid division by zero on bins not visited, the visit count is forced to one on those bins, and the histogram is renormalized. The next N samples drawn from $f_X^{(2)}$ will fall in the tails of the original f_X more often than before, so that visits will tend to be more equally spread across output bins. At convergence we must have $\Theta_{n+1,i} = \Theta_{n,i}$, which from (14) implies $\hat{H}_{n,i}^* = 1/c_n$ for all bins, i.e., a flat histogram (UW-FH).

The MMC update strategy benefits from a general advantage of IS estimators: it provides an unbiased estimate at every cycle, given the weight at the previous cycle, since from (14) we get

$$E[\Theta_{n+1,i}] = E[\hat{H}_{n,i}^*]c_n\Theta_{n,i} = P_i \quad (15)$$

where (13) was used in the second equality. Actually, a bias was introduced on those bins whose occupancy was forced artificially from zero to one.

In the assumption of independent samples, the relative error on estimate $\Theta_{n+1,i}$ on the visited bins is, from (10),

$$\varepsilon_{n+1,i} = \frac{1}{N}\left\{\frac{1}{E[\hat{H}_{n,i}^*]} - 1\right\} = \frac{1}{N}\left\{\frac{c_n\Theta_{n,i}}{P_i} - 1\right\} \quad (16)$$

which from (11) is seen to flatten out for all bins to the value $\frac{M-1}{N}$ at convergence to the UW-FH. Hence, in an ideal setting with independent samples, if the desired SRE on all bins is $\tilde{\varepsilon}$ and we have M bins, the cycle size N should be selected as

$$N \geq \frac{M-1}{\tilde{\varepsilon}}. \quad (17)$$

Note that, starting from any initial guess $\underline{\Theta}_1$, (15) shows that the MMC converges on average even at the first cycle on all visited bins, but with wide fluctuations, i.e., large relative error (16), on those bins in which the probability is largely over-estimated ($\Theta_{n,i} \gg P_i$) and thus the average histogram $E[\hat{H}_{n,i}^*] \to 0$. The usual choice of the uniform distribution for $\underline{\Theta}_1$ makes the relative error at the first steps large in the tail bins, where the histogram count is small. If we have a rough idea of the shape of the PMF $\underline{P}$ to be estimated, a better strategy is to initialize $\underline{\Theta}_1$ to that shape.

B. Smoothed MMC

The stochastic fluctuations due to a finite cycle size N may make $\hat{H}_{n,i}^*$ differ widely from its expectation P^*, hence inducing fluctuations in the estimation even when convergence has already been practically achieved. Some sort of smoothing is thus necessary, as in adaptive equalization [25]. A clever smoothing strategy was suggested by Berg [21]. Consider the logarithm of the ratio of the estimated PMF in adjacent bins; per (14), this quantity follows the update law

$$\beta_{n,i} \triangleq \log\left(\frac{\Theta_{n,i}}{\Theta_{n,i-1}}\right) = \beta_{n-1,i} + \delta_{n,i} \quad (18)$$

978-1-4244-4148-8/09/$25.00 ©2009

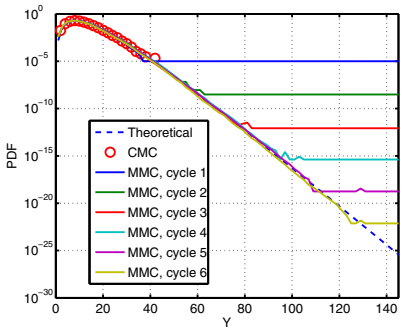

Figure 3. MMC PDF estimate at cycles 1...6 with cycle size 10^5; MC estimate with $6 * 10^5$ samples (circles); true Chi-Square PDF (dashed).

where $\delta_{n,i} \triangleq \log(\hat{H}^*_{n,i}/\hat{H}^*_{n,i-1})$ is a noisy estimate at cycle n of the *log-ratio* of the output PMF in the warped space $\underline{P}^*$. Note that $\beta_{n,i}$ is proportional to the slope at bin i of the estimated output PDF $\Theta_n(y)\Delta y$ plotted versus y in a log scale. This update has the advantage over (14) of exploiting the bin smoothing implicit in using information from two adjacent bins; smoothing over more than two bins has also been proposed [5].

Berg suggests to use the following modified update [21]

$$\beta_{n,i} = \beta_{n-1,i} + \tilde{G}_{n,i}\delta_{n,i} \tag{19}$$

where

$$\tilde{G}_{k,i} = \frac{g_{k,i}}{\sum_{t=1}^{k} g_{t,i}}$$

and

$$g_{t,i} = N\frac{\hat{H}^*_{t,i-1}\hat{H}^*_{t,i}}{\hat{H}^*_{t,i-1} + \hat{H}^*_{t,i}} \tag{20}$$

Reliability factors $\tilde{G}_{k,i}$ are found at time k by normalizing over the samples $g_{t,i}$ available up to time k.

The update law (19) has the classical form found in adaptive equalization, $\delta_{n,i}$ playing the role of the innovation, and $\tilde{G}_{n,i}$ that of the step size, which decreases to zero as the number of time cycles increases and we approach the steady state. Such a decrease provides the desired smoothing in time.

Berg's update (19) can be explicitly rewritten in terms of the original PMFs as the *smoothed MMC update* [3], [21]

$$\frac{\Theta_{n,i}}{\Theta_{n,i-1}} = \frac{\Theta_{n-1,i}}{\Theta_{n-1,i-1}}\left[\frac{\hat{H}^*_{n,i}}{\hat{H}^*_{n,i-1}}\right]^{\tilde{G}_{n,i}}. \tag{21}$$

One advantage of the smoothed MMC update (19) with respect to the original IS update (14) is seen in the case of zero visits. Whenever $\hat{H}^*_{n,i} = 0$ or $\hat{H}^*_{n,i-1} = 0$ the factor $g_{n,i}$ in (20) is zero as is the reliability factor $\tilde{G}_{n,i}$. When both $\hat{H}^*_{n,i}$ and $\hat{H}^*_{n,i-1}$ are zero, we define $g_{n,i} = \tilde{G}_{n,i} = 0$. Hence the smoothed MMC in those cases is evaluated as $\frac{\Theta_{n,i}}{\Theta_{n,i-1}} = \frac{\Theta_{n-1,i}}{\Theta_{n-1,i-1}}$. This causes a propagation of the value of bin $i-1$ to bin i, and it induces a floor (i.e., a bias) in the estimated PMF for those contiguous bins with zero hits in the warped system. As an example, consider estimating the PDF of $Y = \sum_{i=1}^{10} X_i^2$ with X_i independent zero-mean Gaussian RV's with unit variance. Fig. 3 shows the smoothed MMC estimation at cycles 1 to 6, with $N = 10^5$ samples per cycle, along with MC estimation with $6{*}10^5$ samples and the true chi-square PDF (dashed). Here we used 75 bins of width $\Delta y = 2$. From the figure, we see that after 6 cycles the MMC estimate correctly approximates the true PDF down to $\sim 10^{-20}$, while, at the same number of samples, the MC estimate remains at about 10^{-5}, with an MMC gain of 15 orders of magnitude in PDF estimation with respect to MC.

C. Drawing warped samples

The generation of samples from the warped input distributions needed in MMC, which are likely to have a very irregular form in a high dimensional space, is obtained with the very general MCMC method. As explained in the Appendix, a new sample X_t at time t is generated from the sample generated at time $t-1$ and either accepted or rejected based on the odds ratio (24), which only requires the calculation of $g(X_t)$, i.e., a single system evaluation. In this way, samples are generated from the desired $\frac{f_X(x)}{c_n\Theta_n(x)}$ without *a priori* knowledge of the domains D_i in which the input state space gets partitioned by the function $g(\cdot)$. In the Appendix we also point out that sampling from the desired distribution is obtained, i.e., ergodicity is achieved, only when the number of samples per cycle N is sufficiently large. Hence the choice of N may seem critical for a correct sampling. However, in practice for both MMC and other FH algorithms such as WL [24] this is not a key problem; even if the cycle length is not long enough, the next cycles will correct such lack of ergodicity, and explore the state space more evenly. What matters is not correct sampling from the warped PDFs, but convergence to the FH distribution.

MCMC is in widespread use today in statistics and is routinely used in FH algorithms, including MMC. An advantage of the MCMC sample generation method is that the input PDF need only be known up to a multiplicative constant, hence the constant c_n need not be evaluated; this can be a tremendous computational savings for some high dimensional input spaces [21]. A drawback is that samples are correlated, thus making the estimation of the error in the MMC PDF estimation more laborious than with independent samples [8].

IV. Practical aspects of MMC

Acceleration of Monte Carlo techniques is required when the error rate of interest is very low, and/or because the numerical model has a very large number of inputs and/or because the numerical model is computationally intense. In these situations MMC can be of greatest use, as evidenced by its numerous applications in optical communications, where models of optical components may be very time-consuming [2], [17].

Optical systems are not the only ones for which MMC techniques are applicable, although this potential remains largely untapped. Multiuser detection has noise that is inherently non-Gaussian and highly structured. Detector complexity can span from simple linear processing to maximum likelihood sequence estimation. Very accurate numerical models are available, but the computational burden of these models leads to the use of less accurate semi-analytical models to predict behavior. Multi-input, multi-output antenna systems can have significant

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

correlation between signal paths, but this is mostly ignored in analysis. The burgeoning field of cognitive radio requires the optimization of resource allocation across perhaps disparate networks; numerical models can be instrumental in capturing practical channel issues about training sequences and overhead, channel estimation and feedback on channel conditions.

If the SER of interest is very high, on the order of 10^{-3} when forward error correction (FEC) is used, then MMC is not a good accelerator. Other importance sampling techniques such as stratified sampling [26] may be more appropriate, although significant side information about the system under test may be required. MMC is also challenging to use when the system under test includes FEC. The introduction of FEC leads to isolated islands in the input space being responsible for error events. With isolated islands, the MCMC exploration of critical regions of the input space can be difficult ([27], Ch. 31). Nonetheless some researchers have partially succeeded in using MMC to test numerical models with FEC [28], [29]. Note that these deficiencies are not unique to MMC; indeed all Monte Carlo techniques have difficulty exploring FEC performance.

A. Input Vector Correlations

As one can easily understand from the Appendix on MCMC, one problem of the state space exploration with a symmetric Metropolis candidate chain is that no preferential directions are present in the exploration. Hence such a method is most effective in sampling input distributions f_X with independent components, while lower efficiency is obtained when correlations are present [27]. In such a case more sophisticated exploration criteria such as Hamiltonian and related methods should be used ([27], Ch. 30).

There is, however, a countermeasure for correlations for most non-pathological cases. As long as the input process is wide sense stationary, we are assured by Wold's decomposition theorem [30] that a whitening filter exists. Such a filter can be included as part of the system, and an input distribution with uncorrelated components can be used. The whitening operation is quite effective in dealing with Gaussian vectors, since lack of correlation implies independence. The trade-off here is clearly the analytical pre-calculation of the whitening filter.

This issue is closely related to the scaling of the simulation time with the dimension of the input vector X. Although in MCMC the state space can be continuous, thinking of such a space as discrete and recalling the MCMC random walk in state space described in the Appendix helps us develop intuition about the scaling rule.

Suppose the dimension of the input state is K, and b_x is the number of states per input random component and that this provides adequate resolution for the simulation. For the case of dependent components in X we must create a K-dimensional input space and test all possible combinations of the ordered pairs in generating samples according to our warped distribution. Hence the input PDF spans a K-dimensional space and we require b_x^K states, i.e. an exponential increase with K in the number of states in the Markov chain. If the components are instead independent, we just need to correctly sample each of them on b_x states, hence the exploration complexity scales linearly with K.

B. Dealing with system memory

Consider a system with memory where the state of the system during M previous symbol intervals will affect the system output. For MMC simulations, the input vector must be long enough to include all inputs that fall within that memory range.

In systems with memory (whether arising from filtering or other sources), the memory is often captured in numerical models by use of a shift-register model; register length must be sufficient to accommodate an input vector capturing all memory effects. Two simulation approaches are possible. In a sequential approach, the register contents are passed to the system under test, an output symbol is generated, and a new input is shifted into the register. A single new input is generated at each time step. In a non-sequential approach, the register contents are passed to the system under test, an output symbol is generated, and an entirely new block of data enters the system at the next time step. If the register is long enough to include all system inputs affecting the output symbol, the non-sequential system gives a time down-sampled version of the sequential system output, and if the output processes are stationary both techniques will yield accurate statistics. However, the block simulation method has significantly faster convergence of MCMC space exploration as compared with the sequential approach.

C. Discretization of the output space

The choice of bin width Δy which defines the bins B_i in the output space is critical for proper operation of MMC. If Δy is too small, a very high number of samples is required for an accurate estimate of the output PMF $\Theta_{n,i}$. If, on the other hand, Δy is too large, we may encounter very large deviations in the PMF for two adjacent bins B_i and B_{i+1}: $\Theta_{n,i} \gg \Theta_{n,i+1}$. In such a case, the odds ratio of (26) would be very small, and the MCMC machine will move too slowly in the exploration of the state space. We empirically find that the bin width should be chosen such that adjacent bins have probabilities within one order of magnitude of one other.

D. Exploration of the input space

As shown in (27) of the Appendix, the MCMC machine needs a vector U to produce a future state X of the chain. If the components of X are independent and identically distributed (i.i.d.), then the components of U are i.i.d. uniform random variables. The k^{th} element of U is denoted by U_k, and is distributed over the range $[-\Delta_U/2, +\Delta_U/2]$. The value of Δ_U is a key parameter for the MCMC algorithm to correctly sample the input space. Intuitively, if it were too big then the proposed state would likely fall very far from the present state. This would lead to a high rejection ratio, and hence the chain would hardly move. On the other hand, if Δ_U were too small, the rejection ratio would be higher but the steps would be very small, hence the chain would move very slowly and it would take a very high number of samples for it to reach the steady state. We empirically find that a good compromise is $\Delta_U \sim \sigma$, where σ is the standard deviation of the known true distribution of the i.i.d. components of the input vector.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

E. Choice of number of cycles vs. samples per cycle

We end with a final note on the best cycle size in MMC. For N_{cycle} cycles and N samples per cycle, we will generate a total of $N_{cycle}N$ samples. We empirically find that, in order to resolve the estimated PDF down to a desired level, the product $N_{cycle} * log(K_1/(N - N_0))$ must remain constant, with K_1 a suitable constant, and with $N > N_0$, i.e., with a cycle size lower bounded by a value N_0 related to the bound in (17). The optimal N (along the above constraint curve) that minimizes the total cost $N_{cycle}N$ is usually close to the lower bound N_0. The message here is that it is not necessary to make N very large (e.g. in order to achieve ergodicity in the sampling MCMC), but a smaller cycle size and more cycles achieve the same goal at a lower total cost. Unfortunately, N_0 is widely problem-dependent, so that in practice the cycle size gets selected by trial and error: it is typically larger for a smaller desired PDF level to be resolved.

V. Conclusion

We have presented guidelines on the exploitation of the MMC adaptive importance sampling technique for the analysis of symbol error rate performance of communications systems. Salient features of the MMC adaptation were described, to better prepare researchers to mold their simulation environments to that of MMC. The good convergence and efficient computation of MMC requires that care be taken in casting the noise and other inputs in an appropriate manner. We discussed the importance of incorporating into the system under test mechanisms to correctly correlate inputs. We discussed the advantages of block based simulations over sequential simulations. We provided several rules of thumb for tweaking various MMC parameters.

Appendix: MCMC

MCMC fundamentals

MCMC is a technique to produce samples from a desired, analytically known probability density function $f_X(x)$, with X taking values in a multidimensional space Γ. Without loss of generality, and for the sake of clarity, we consider a discretized space Γ [31], i.e. we have a known PMF $\underline{p}_X = [p_X(x_1), p_X(x_2), \ldots]$, with $p_X(x_i) \cong f_X(x_i)\Delta x$, for the discretized states $\{x_i\}_{i=1}^{\infty}$ in Γ. MCMC synthesizes the desired samples $\{X_m,\, m \geq 1\}$ from a memoryless sequence, i.e., a discrete-time Markov Chain (DTMC) whose steady-state distribution $\underline{\pi}$ coincides with the desired PMF $\underline{p}_X$.

A DTMC is characterized by its transition matrix $\mathbf{P} = \{p_{ij}\}$, with transition probability from any state x_i to any state x_j defined as $p_{ij} = P\{X_m = x_j \,|\, X_{m-1} = x_i\}$. The steady-state distribution solves the equation [32]

$$\underline{\pi} = \underline{\pi}\mathbf{P}\,. \tag{22}$$

While the classical DTMC problem is to find $\underline{\pi}$ for a given $\mathbf{P}$, the MCMC problem is conversely to find a matrix $\mathbf{P}$ which satisfies (22) for a known $\underline{\pi} \triangleq \underline{p}_X$. We clearly require the DTMC to be ergodic, i.e., that $\mathbf{P}$ has a unique $\underline{\pi}$, and that the PMF of the chain at time m, namely $\underline{p}(m) = [P\{X_m = x_1\}, P\{X_m = x_1\}, \ldots]$, converges to $\underline{\pi}$ as $m \to \infty$. Thus the shortcomings of the MCMC method are that

i) the sequence $\{X_m,\, m \geq 1\}$ will reflect the desired limiting ditribution $\underline{p}_X$ only for large enough m, and

ii) the samples will be correlated according the random walk on the states driven by the matrix $\mathbf{P}$.

There are clearly infinitely many ergodic matrices $\mathbf{P}$ that solve (22), and we need just one. A unique, simple solution is found by imposing the extra constraint that the DTMC be time-reversible. A necessary and sufficient condition for time reversibility is that, at steady-state, for every pair of states (x_i, x_j) the probability of being at x_i at time $m-1$ and moving to x_j at time m equals the probability of being at x_j at $m-1$ and moving to x_i at m [32]

$$\pi_i p_{ij} = \pi_j p_{ji}\,. \tag{23}$$

These are called local balance equations and they determine all the unknowns $\{p_{ij}\}$.

A clever way of practically implementing a reversible DTMC with this method was introduced by Metropolis [19] in 1953 and 17 years later generalized by Hastings [20]. Hastings proposed the following procedure to find the $\{p_{ij}\}$

i) Start with any transition matrix $\mathbf{Q} = \{q_{ij}\}$, called the *candidate* chain;

ii) for any pair of states x_i, x_j, $i \neq j$, which do not satisfy (23) a randomization procedure is introduced such that every time the candidate chain proposes a move $i \to j$ the move is accepted with probability α_{ij} and otherwise rejected (i.e. the chain remains in the same state at the next time). Hence $p_{ij} = \alpha_{ij} q_{ij}$.

For arbitrary choice of Q, it may happen that either (a) $\pi_i q_{ij} > \pi_j q_{ji}$ or (b) $\pi_i q_{ij} < \pi_j q_{ji}$. In case (a) we accept all transitions $j \to i$, i.e. use $\alpha_{ji} = 1$ (hence $p_{ji} = q_{ji}$), and decrease the transitions $i \to j$ by accepting a fraction $\alpha_{ij} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}} < 1$ of such moves so as to reach equality as in (23). In case (b) we swap the roles of i and j, so that in general $\alpha_{ij} = \min[1, R_{ij}]$, where

$$R_{ij} = \frac{\pi_j q_{ji}}{\pi_i q_{ij}} = \frac{f_X(x_j) q_{ji}}{f_X(x_i) q_{ij}} \tag{24}$$

is the *odds ratio*, and we have substituted back the original PDF of the input RV X. Note that, since only the ratio of PDFs at the two states is needed, such a PDF need only be known up to a normalization constant. There is no need to normalize the PDF to generate samples from it. In some physical settings the normalization constant is impractical or impossible to compute [21] and the MCMC algorithm offers the only known solution to this simulation problem.

Metropolis MCMC [19] uses a symmetric candidate $q_{ij} = q_{ji}$ so that the the odds ratio further simplifies. Starting from initial state x_i, common practice is to select the Metropolis candidate as $x_j = x_i + U$ where U is a uniform random vector in space Γ. No quantization is needed in the input space. The variance of U is important in determining both the acceptance ratio and the speed of exploration of the chain in the input space, and is one of the key tuning parameters of the MCMC machine.

Use of MCMC in the MMC algorithm

When generating warped samples at the n-th cycle in an MMC algorithm using the MCMC machine, the odds ratio (24)

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

for the desired UW warping (12) becomes

$$R_{ij} = \frac{\Theta_n(x_i) f_X(x_j) q_{ji}}{\Theta_n(x_j) f_X(x_i) q_{ij}} \tag{25}$$

and the constant c_n cancels out. As suggested in [3], the odds ratio can be simplified to

$$R_{ij} = \frac{\Theta_n(x_i)}{\Theta_n(x_j)} \tag{26}$$

by choosing $q_{ij} = f_X(x_j)\Delta x$, i.e., by having a candidate chain whose transition probability only depends on the final state x_j; the proposed candidate x_j is drawn from the original distribution f_x independently of the initial state x_i. This is known as an independence chain [31]. To find (26), we need only calculate $y_j = g(x_j)$ for the selected candidate x_j ($y_i = g(x_i)$ was already calculated at the previous sample) to determine to which bin it belongs and thus determine the value of $\Theta_n(x_j)$, i.e., the intermediate estimate of the output PMF at cycle n of such a bin.

A direct use of the candidate independence chain would clearly lead to too many rejections in a large K-dimensional state space Γ. Hence in [3] it is suggested to implement the candidate chain itself using an MCMC machine with component-wise independent Metropolis reject/accept mechanisms: this technique is known as concatenation [27] or one-variable-at-a-time [31], and works as follows. For all components $1 \leq k \leq K$

i) starting from the k-th component $x_{k,i}$ of vector x_i the k-th component of candidate vector x_j is Metropolis generated as

$$x_{k,j} = x_{k,i} + U_k \tag{27}$$

with U_k a scalar uniform RV;

ii) if $G_k(\cdot)$ is the marginal PDF of $f_x(\cdot)$ for the k-th component of vector x, the move $x_{k,i} \rightarrow x_{k,j}$ is accepted for the candidate with probability $\alpha_{ij}^{(k)} = \min[1, \frac{G_k(x_{k,j})}{G_k(x_{k,i})}]$; if the move is rejected, $x_{k,j} = x_{k,i}$.

It can be shown that if X has independent components, i.e., $f_x(x_i) = \Pi_{i=1}^{K} G_k(x_{k,i})$, then $\frac{q_{ji}}{q_{ij}} = \frac{f_x(x_i)}{f_x(x_j)}$, and (25) simplifies to (26). Once the new candidate x_j is formed as described previously, the global move $x_i \rightarrow x_j$ is accepted based on the odds ratio (26). Since candidate moves $x_i \rightarrow x_j$ are made at smaller distances by suitable choice of the variance of the Metropolis RVs $\{U_k\}$, the rejection ratios can be substantially decreased, accelerating the state exploration.

REFERENCES

[1] B. A. Berg, and T. Neuhaus, "Multicanonical algorithms for first-order phase transitions," *Phys. Lett. B*, vol. 267, no. 2, pp. 249-253, Sept. 1991.

[2] D. Yevick, "Multicanonical communication system modeling-application to PMD statistics," *IEEE Photon. Technol. Lett.*, vol. 14, no. 11, pp. 1512-1514, 2002.

[3] R. Holzlohner, and C. R. Menyuk, "Use of multicanonical monte carlo simulations to obtain accurate bit error rates in optical communications systems," *Opt. Lett.*, vol. 28, no. 20, pp. 1894-1896, Oct. 2003.

[4] T. Kamalakis, D. Varoutas, and T. Sphicopoulos, "Statistical study of in-band crosstalk noise using the multicanonical Monte Carlo method," *IEEE Photon. Technol. Lett.*, vol 16, no. 10, pp. 2242-2244, 2004.

[5] T. Lu, and D. Yevick, "Efficient multicanonical algorithms," *Photon. Technol. Lett.*, vol. 17, no. 4, pp. 861-863, Apr. 2005.

[6] G. Biondini and W.L. Kath, "Polarization-dependent chromatic dispersion and its impact on return-to-zero transmission formats," *IEEE Photon. Technol. Lett.*, vol. 17, no. 9, pp. 1866-1868, 2005.

[7] A. O. Lima, C. R. Menyuk, and I. T. lima, "Comparing two biasing monte carlo methods for calculating outage probabilities in sytems with multisection PMD compensators," *IEEE Photon. Technol. Lett.*, vol. 17, no. 12, pp. 2580-2582, 2005.

[8] A. O. Lima, I. T. Lima, and C. R. Menyuk, "Error estimation in multicanonical monte carlo simulations with applications to polarization mode dispersion emulators," *J. Lightw. Technol.*, vol. 23, no. 11, Nov. 2005.

[9] W. Pellegrini, J. Zweck, C.R. Menyuk and R. Holzlohner, "Computation of bit error ratios for a dense WDM system using the noise covariance matrix and multicanonical Monte Carlo methods," *IEEE Photon. Technol. Lett.*, vol. 17, no. 8, pp. 1644-1646, 2005.

[10] A. Bilenca and G. Eisenstein, "Statistical noise properties of an optical pulse propagating in a nonlinear semiconductor optical amplifier," *IEEE J. Quantum Electron.*, vol. 41, no. 1, pp. 36-44, 2005.

[11] Y. Yadin and M. Shtaif and M. Orenstein, "Bit-error rate of optical DPSK in fiber systems by multicanonical Monte Carlo Simulations," *IEEE Photon. Technol. Lett.*, vol. 17, no. 6, pp. 1355-1357, 2005.

[12] M. Nazarathy and E. Simony and Y. Yadin, "Analytical evaluation of bit error rates for hard detection of optical differential phase amplitude shift keying (DPASK)," *J. Lightw. Technol.*, vol. 24, no. 5, pp. 2248-2260, 2006.

[13] I. Nasieva, A. Kaliazin, S. K. Turitsyn, "Multicanonical Monte Carlo modelling of BER penalty in transmission systems with optical regeneration," *Opt. Commun.*, vol. 262, pp. 246-249, 2006.

[14] L. Gerardi, M. Secondini, E. Forestieri, "Pattern Perturbation Method for Multicanonical Monte Carlo Simulations in Optical Communications", *IEEE Photon. Technol. Lett.*, vol. 19, pp. 1934-1936, Dec. 2007.

[15] T. I. Lakoba , "Multicanonical Monte Carlo Study of the BER of an All-Optically 2R Regenerated Signal," *IEEE J. Sel. Topics Quantum Electron.*, vol. 14, pp. 599-609, May/June 2008.

[16] A. Ghazisaeidi, F. Vacondio, A. Bononi, and L. A. Rusch, "Statistical Characterization of Bit Patterning in SOAs: BER Prediction and Experimental Validation," in *Proc. OFC 2009*, paper OWE7, San Diego, CA, March 2009.

[17] A. Ghazisaeidi, F. Vacondio, A. Bononi, and L. A. Rusch, "SOA Intensity Noise Suppression: Multicanonical Monte Carlo Simulator of Extremely Low BER," *IEEE J. Lightwave Technol.*, vol. 27, pp. 2667-2677, July 2009.

[18] M. Jeruchim, "Techniques for estimating the bit error rate in the simulation of digital communication systems," *IEEE J. Sel. Areas. Commun.*, vol. SAC-2, pp. 153-170, Jan. 1994.

[19] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no.6, pp. 1087-1092, June 1953.

[20] W. K. Hastings, "Monte Carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, pp. 97-109, Apr. 1970.

[21] B. A. Berg, "Introduction to multicanonical monte carlo simulations," *Fields Instr. Commun.*, vol. 26, pp. 1-24, 2000.

[22] F. Liang, "A theory on flat histogram monte carlo algorithms," *J. Stat. Phys.*, vol. 122, pp. 511-529, Feb. 2006.

[23] Y. F. Atchade, J. S. Liu, "The Wang-Landau algorithm for MC computation in general state spaces", Technical report, University of Ottawa, 2004 (available at www.mathstat.uottawa.ca/~yatch436/gwl.pdf)

[24] F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states" *Phys. Rev. Lett.*, vol. 86, pp. 2050-2053, Mar. 2001.

[25] S. Haykin, *Adaptive Filter Theory*, 4th Ed. Prentice Hall: Englewood Cliffs, NJ, 2001.

[26] P. Serena, N. Rossi, M. Bertolini, and A. Bononi, "Stratified Sampling Monte Carlo Algorithm for Efficient BER estimation in Long-Haul Optical Transmission Systems," *IEEE J. Lightwave Technol.*, vol. 27, pp. 2404-2411, July 2009.

[27] D. J. C. MacKay, "Information Theory, Inference, and Learning Algorithms". Cambridge University Press. 2003.

[28] Y. Iba and K. Hukushima "Testing Error Correcting Codes by Multicanonical Sampling of Rare Events," *J. Phys. Soc. Jpn.*, vol. 77, no. 10, 2008

[29] R. Holzlohner *et al.*, "Evaluation of very low BER of FEC codes using dual adaptive importance sampling," *IEEE Photon. Technol. Lett.*, vol. 9, pp. 163-165, 2005.

[30] A. Papoulis *Probability, Random Variables, and Stochastic Processes*, 3rd Ed. McGraw-Hill: New York, 1991

[31] C. J. Geyer, "Markov Chain Monte Carlo lecture notes" Course notes, Spring Quarter 1998, University of Minnesota.

[32] S. M. Ross, *Stochastic Processes*. Wiley: New York, 1983.

978-1-4244-4148-8/09/$25.00 ©2009

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.